Measuring the Effects of Starting Pitching

Lee Przybylski

Iowa State University

November 20, 2021

Outline

1. Motivation
2. Generalized Linear Mixed Effects Models
3. Model Selection
4. Predictive Value for Game Outcomes
5. Pitching Metrics
6. Closing Thoughts

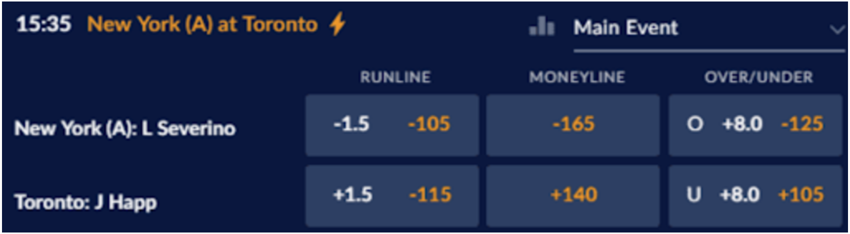
Outline

Github: https://github.com/przybylee/RunsScoredAnalysis

Motivation

A sports book presents the moneyline with the team names and the probable starting pitchers.

A sports book presents the moneyline with the team names and the probable starting pitchers.



Questions

Can we use this information to make a model that uses this information to predict the outcome of future games?

Questions

Can we use this information to make a model that uses this information to predict the outcome of future games?

Some of it, yes

Questions

Can we use this information to make a model that uses this information to predict the outcome of future games?

Some of it, yes

Can we use predictors from this model to gain a better understanding of teams, venues, and starting pitchers?

Questions

Can we use this information to make a model that uses this information to predict the outcome of future games?

Some of it, yes

Can we use predictors from this model to gain a better understanding of teams, venues, and starting pitchers?

Yes

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

$$log(\lambda_{ijkl}) = \mu + \omega_i + \delta_j + \nu_k + \chi \mathbf{1}_{ik}. \tag{1}$$

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

$$log(\lambda_{ijkl}) = \mu + \omega_i + \delta_j + \nu_k + \chi \mathbf{1}_{ik}. \quad (1)$$

We let $\mathbf{1}_{ik} = 1$ if team i is playing at home in game k and $\mathbf{1}_{ik} = 0$ otherwise.

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

$$log(\lambda_{ijkl}) = \mu + \omega_i + \delta_j + \nu_k + \chi \mathbf{1}_{ik}. \quad (1)$$

We let $\mathbf{1}_{ik} = 1$ if team i is playing at home in game k and $\mathbf{1}_{ik} = 0$ otherwise.

We will refer to this as the *generalized linear model* (GLM).

Model Selection

We use fit the GLM in R using data from 2021 and we get some noteworthy estimates:

Model Selection

We use fit the GLM in R using data from 2021 and we get some noteworthy estimates:

- $\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ} = 1.192$, meaning the expected runs scored by Houston batting in Arizona is $\exp(\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ}) \approx 6.85$

Model Selection

We use fit the GLM in R using data from 2021 and we get some noteworthy estimates:

- $\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ} = 1.192$, meaning the expected runs scored by Houston batting in Arizona is $\exp(\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ}) \approx 6.85$
- $\hat{\nu}_{COL} - \hat{\nu}_{ARZ} \approx 0.15$, meaning teams score an average of $e^{0.15} - 1 \approx 16\%$ more runs at Coors Field.

Model Selection

We use fit the GLM in R using data from 2021 and we get some noteworthy estimates:

- $\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ} = 1.192$, meaning the expected runs scored by Houston batting in Arizona is $\exp(\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ}) \approx 6.85$
- $\hat{\nu}_{COL} - \hat{\nu}_{ARZ} \approx 0.15$, meaning teams score an average of $e^{0.15} - 1 \approx 16\%$ more runs at Coors Field.
- $\hat{\chi} \approx 0.04$, meaning a team scores about 3% more runs when playing at home.

Model Selection

We use fit the GLM in R using data from 2021 and we get some noteworthy estimates:

- $\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ} = 1.192$, meaning the expected runs scored by Houston batting in Arizona is $\exp(\hat{\mu} + \hat{\omega}_{HOU} + \hat{\delta}_{ARZ} + \hat{\nu}_{ARZ}) \approx 6.85$
- $\hat{\nu}_{COL} - \hat{\nu}_{ARZ} \approx 0.15$, meaning teams score an average of $e^{0.15} - 1 \approx 16\%$ more runs at Coors Field.
- $\hat{\chi} \approx 0.04$, meaning a team scores about 3% more runs when playing at home.
- The Null deviance is 43853 on 4922 degrees of freedom.

Model Selection

In order to combine information to make better predictions about team, defense, and venue effects, we will adopt a generalized linear mixed effects (GLMM) model.

Model Selection

In order to combine information to make better predictions about team, defense, and venue effects, we will adopt a generalized linear mixed effects (GLMM) model.

We let $y_{ijkl} \sim \text{Poisson}(\lambda_{ijklm})$ be the number of runs scored by team i against team j at venue k, during game l. We have

$$\log(\lambda_{ijkl}) = \mu + \chi \mathbf{1}_{il} + b_i + f_j + v_k, \tag{2}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2),$$

Model Selection

In order to combine information to make better predictions about team, defense, and venue effects, we will adopt a generalized linear mixed effects (GLMM) model.

We let $y_{ijkl} \sim \text{Poisson}(\lambda_{ijklm})$ be the number of runs scored by team i against team j at venue k, during game l. We have

$$\log(\lambda_{ijkl}) = \mu + \chi \mathbf{1}_{il} + b_i + f_j + v_k, \tag{2}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2),$$

We will refer to this as the *reduced GLMM* (R. GLMM)

Model Selection

We use fit the R.GLMM and see the following estimates:

Model Selection

We use fit the R.GLMM and see the following estimates:

- $\hat{\mu} = 1.47$, meaning the expected runs for the visiting team is 4.38.

Model Selection

We use fit the R.GLMM and see the following estimates:

- $\hat{\mu} = 1.47$, meaning the expected runs for the visiting team is 4.38.
- $\hat{\chi} = 0.036$, meaning a team scores 3.8% more runs when playing at home.

Model Selection

We use fit the R.GLMM and see the following estimates:

- $\hat{\mu} = 1.47$, meaning the expected runs for the visiting team is 4.38.
- $\hat{\chi} = 0.036$, meaning a team scores 3.8% more runs when playing at home.
- $\hat{\sigma}_b^2 = 0.008$, (Variance in offense)

Model Selection

We use fit the R.GLMM and see the following estimates:

- $\hat{\mu} = 1.47$, meaning the expected runs for the visiting team is 4.38.
- $\hat{\chi} = 0.036$, meaning a team scores 3.8% more runs when playing at home.
- $\hat{\sigma}_b^2 = 0.008$, (Variance in offense)
- $\hat{\sigma}_f^2 = 0.012$, (Variance in defense)

Model Selection

We use fit the R.GLMM and see the following estimates:

- $\hat{\mu} = 1.47$, meaning the expected runs for the visiting team is 4.38.
- $\hat{\chi} = 0.036$, meaning a team scores 3.8% more runs when playing at home.
- $\hat{\sigma}_b^2 = 0.008$, (Variance in offense)
- $\hat{\sigma}_f^2 = 0.012$, (Variance in defense)
- $\hat{\sigma}_v^2 = 0.010$ (Variance in park effects)

Model Selection

Because this model is likely to have overdispersion, we add random effects for each game and each team within a game.

Model Selection

Because this model is likely to have overdispersion, we add random effects for each game and each team within a game.

We let $y_{ijklm} \sim \text{Poisson}(\lambda_{ijklm})$ be the number of runs scored by team i against team j at venue k, against starting pitcher l during game m. We have

$$\log(\lambda_{ijklm}) = \mu + \chi \mathbf{1}_{im} + b_i + f_j + v_k + p_l + g_m + e_{im}, \tag{3}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2),$$

$$g_m \overset{\text{iid}}{\sim} N(0, \sigma_g^2), e_{im} \overset{\text{iid}}{\sim} N(0, \sigma_e^2).$$

Model Selection

Because this model is likely to have overdispersion, we add random effects for each game and each team within a game.

We let $y_{ijklm} \sim \text{Poisson}(\lambda_{ijklm})$ be the number of runs scored by team i against team j at venue k, against starting pitcher l during game m. We have

$$\log(\lambda_{ijklm}) = \mu + \chi \mathbf{1}_{im} + b_i + f_j + v_k + p_l + g_m + e_{im}, \tag{3}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2),$$

$$g_m \overset{\text{iid}}{\sim} N(0, \sigma_g^2), e_{im} \overset{\text{iid}}{\sim} N(0, \sigma_e^2).$$

We will refer to this as the *full GLMM* (F. GLMM)

Model Selection

We use fit the F.GLMM and see the following estimates:

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$
- $\hat{\sigma}_v^2 = 0.007$

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$
- $\hat{\sigma}_v^2 = 0.007$
- $\hat{\sigma}_p^2 = 0.002$ (Variance in starting pitchers)

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$
- $\hat{\sigma}_v^2 = 0.007$
- $\hat{\sigma}_p^2 = 0.002$ (Variance in starting pitchers)
- $\hat{\sigma}_g^2 = 0.006$ (Variance in game effects)

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$
- $\hat{\sigma}_v^2 = 0.007$
- $\hat{\sigma}_p^2 = 0.002$ (Variance in starting pitchers)
- $\hat{\sigma}_g^2 = 0.006$ (Variance in game effects)
- $\hat{\sigma}_e^2 = 0.229$ (Overdispersion)

Model Selection

We use fit the F.GLMM and see the following estimates:

- $\hat{\mu} = 1.36$, meaning the expected runs for the visiting team is 3.91.
- $\hat{\chi} = 0.036$, meaning teams tend to score 3.8% more runs at home.
- $\hat{\sigma}_b^2 = 0.005$
- $\hat{\sigma}_f^2 = 0.010$
- $\hat{\sigma}_v^2 = 0.007$
- $\hat{\sigma}_p^2 = 0.002$ (Variance in starting pitchers)
- $\hat{\sigma}_g^2 = 0.006$ (Variance in game effects)
- $\hat{\sigma}_e^2 = 0.229$ (Overdispersion)

A large estimate for σ_e^2.

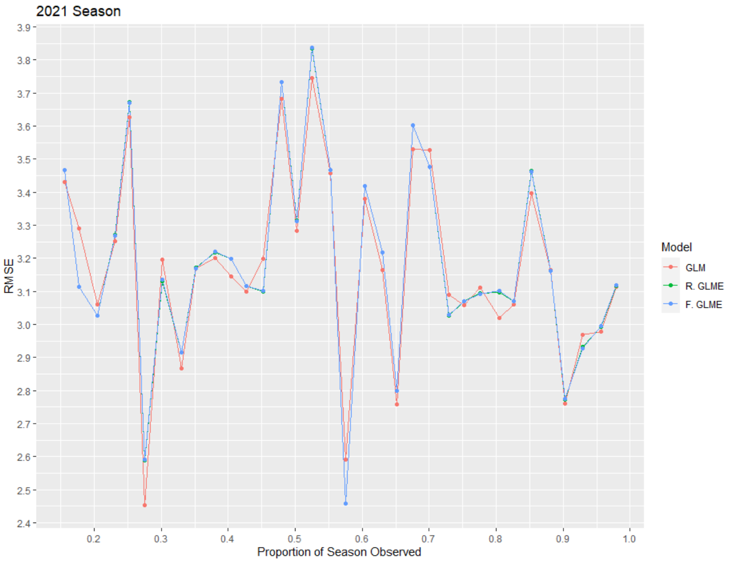
Predictive Value for Game Outcomes

A likelihood ratio test shows that the full GLME model is a much better fit than the reduced version.

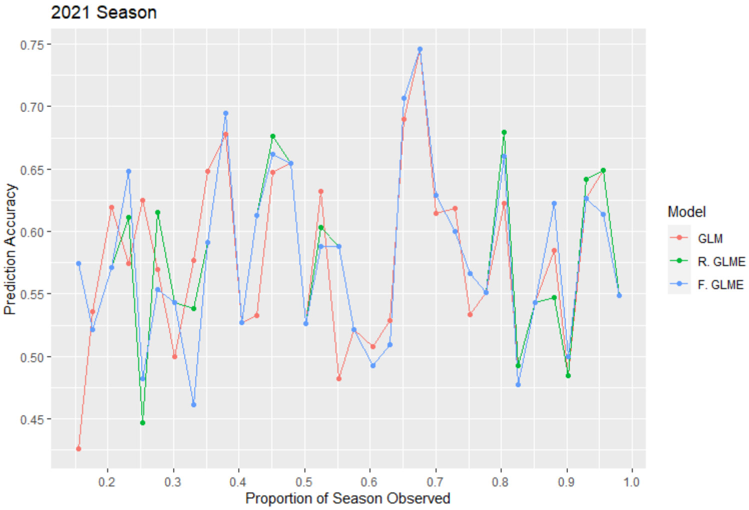
To understand the predictive power of these models, we train the model using results from the first 15% of the season and predict on the next 2.5% of the season. This is about 110 games.

We measure the accuracy of our predictions, then retrain using the first 17.5% of the season and predict on the next 2.5% of the season. We measure and repeat this process for the rest of the year.

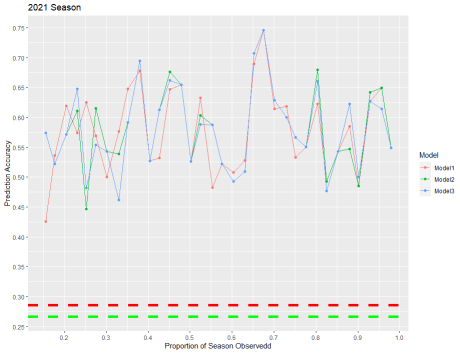
Predictive Value for Game Outcomes



Model Testing



Predictive Value for Game Outcomes



Surprisingly the accuracy of the opening and closing money lines was very low. To measure this, we took the proportion of times the losing team had a positive moneyline.

Pitching Metrics

We can extract the BLUPs for the p_j's and assign a value to each pitcher. For each pitcher, we will report $e^{\hat{p}_j}$ since this is the proportion of runs the pitcher would allow compared to the average pitcher in his environment. We call this score the Starting Pitcher Rating (SRP).

Pitching Metrics

We can extract the BLUPs for the p_j's and assign a value to each pitcher. For each pitcher, we will report $e^{\hat{p}_j}$ since this is the proportion of runs the pitcher would allow compared to the average pitcher in his environment. We call this score the Starting Pitcher Rating (SRP).

Example: Gerit Cole has an SPR of 0.989 for the 2021 season. This suggests that when he is starting, his team allows about 1% less runs than with the average pitcher in his situation.

Pitching Metrics

We can extract the BLUPs for the p_j's and assign a value to each pitcher. For each pitcher, we will report $e^{\hat{p}_j}$ since this is the proportion of runs the pitcher would allow compared to the average pitcher in his environment. We call this score the Starting Pitcher Rating (SRP).

Example: Gerit Cole has an SPR of 0.989 for the 2021 season. This suggests that when he is starting, his team allows about 1% less runs than with the average pitcher in his situation.

We consider correlation of SPR with other pitching metrics (ERA, FIP, DRA, WARP).

Pitching Metrics

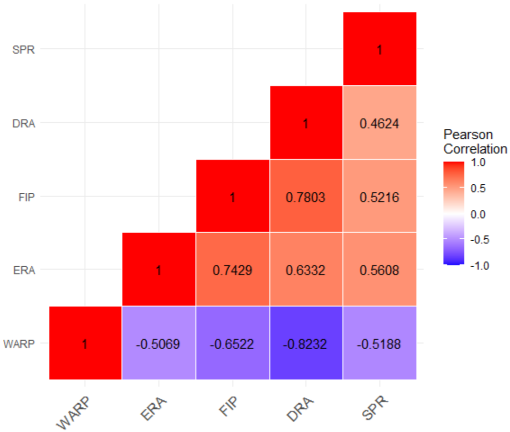


Figure: Correlation Between Metrics in 2021

Pitching Metrics

Another value in a pitching metric is its predictive value for future seasons. This has been a major goal since the research by Voros McCracken in the early 2000's that lead us to FIP.

Pitching Metrics

Another value in a pitching metric is its predictive value for future seasons. This has been a major goal since the research by Voros McCracken in the early 2000's that lead us to FIP.

	2015-	2016-	2017-	2018-
#	140	138	151	163
WARP	0.624	0.560	0.549	0.474
DRA	0.562	0.604	0.546	0.518
FIP	0.421	0.426	0.353	0.406
ERA	0.226	0.237	0.228	0.119
SRP	0.158	0.153	0.103	-0.017

Table: Correlation Across Consecutive Seasons

ERA Leaders

rank	Name	Team	ERA	SPR
1	Aaron Loup	NYM	0.950	0.998
2	Jacob deGrom	NYM	1.080	0.970
3	Dominic Leone	SFO	1.510	0.999
4	Collin McHugh	TAM	1.550	0.989
5	Jesse Chavez	ATL	2.140	0.994
6	Tyler Rogers	SFO	2.220	0.981
7	Louis Head	TAM	2.310	0.999
8	Drew Smith	NYM	2.400	1.008
9	Corbin Burnes	MIL	2.430	0.963
10	Ryan Burr	CWS	2.450	1.001

FIP Leaders

rank	Name	Team	FIP	SPR
1	Jacob deGrom	NYM	1.230	0.970
2	Corbin Burnes	MIL	1.630	0.963
3	Jesse Chavez	ATL	2.010	0.994
4	Collin McHugh	TAM	2.120	0.989
5	Taylor Rogers	MIN	2.130	0.981
6	Aaron Loup	NYM	2.440	0.998
7	Trevor Rogers	MIA	2.540	0.981
8	Tanner Houck	BOS	2.570	0.996
9	Zack Wheeler	PHI	2.590	0.961
10	Logan Webb	SFO	2.720	0.973

DRA Leaders

rank	Name	Team	DRA	SPR
1	Jacob deGrom	NYM	2.410	0.970
2	Corbin Burnes	MIL	2.630	0.963
3	Taylor Rogers	MIN	3.000	0.981
4	Michael Kopech	CWS	3.040	0.994
5	Tyler Glasnow	TAM	3.070	0.980
6	Zack Wheeler	PHI	3.150	0.961
7	Brandon Woodruff	MIL	3.180	0.993
8	Gerrit Cole	NYY	3.250	0.989
9	Max Scherzer		3.260	0.953
10	Logan Webb	SFO	3.290	0.973

2021 SPR Leaders

rank	Name	Team	SPR	ERA	FIP	WARP	DRA
1	Max Scherzer		0.953	2.460	2.970	4.700	3.260
2	Zack Wheeler	PHI	0.961	2.780	2.590	5.800	3.150
3	Corbin Burnes	MIL	0.963	2.430	1.630	5.500	2.630
4	Shane Bieber	CLE	0.968	3.170	3.020	2.400	3.350
5	Jacob deGrom	NYM	0.970	1.080	1.230	3.300	2.410
6	Logan Webb	SFO	0.973	3.030	2.720	3.600	3.290
7	Lance Lynn	CWS	0.977	2.690	3.310	3.100	3.840
8	Chris Flexen	SEA	0.977	3.610	3.890	0.700	5.220
9	Blake Snell	SDG	0.978	4.200	3.820	2.400	3.930
10	Robbie Ray	TOR	0.979	2.840	3.690	3.900	3.760

2021 SPR Leaders

rank	Name	Team	SPR	ERA	FIP	WARP	DRA
1	Max Scherzer		0.953	2.460	2.970	4.700	3.260
2	Zack Wheeler	PHI	0.961	2.780	2.590	5.800	3.150
3	Corbin Burnes	MIL	0.963	2.430	1.630	5.500	2.630
4	Shane Bieber	CLE	0.968	3.170	3.020	2.400	3.350
5	Jacob deGrom	NYM	0.970	1.080	1.230	3.300	2.410
6	Logan Webb	SFO	0.973	3.030	2.720	3.600	3.290
7	Lance Lynn	CWS	0.977	2.690	3.310	3.100	3.840
8	Chris Flexen	SEA	0.977	3.610	3.890	0.700	5.220
9	Blake Snell	SDG	0.978	4.200	3.820	2.400	3.930
10	Robbie Ray	TOR	0.979	2.840	3.690	3.900	3.760

Who?

2021 SPR Leaders

rank	Name	Team	SPR	ERA	FIP	WARP	DRA
1	Max Scherzer		0.953	2.460	2.970	4.700	3.260
2	Zack Wheeler	PHI	0.961	2.780	2.590	5.800	3.150
3	Corbin Burnes	MIL	0.963	2.430	1.630	5.500	2.630
4	Shane Bieber	CLE	0.968	3.170	3.020	2.400	3.350
5	Jacob deGrom	NYM	0.970	1.080	1.230	3.300	2.410
6	Logan Webb	SFO	0.973	3.030	2.720	3.600	3.290
7	Lance Lynn	CWS	0.977	2.690	3.310	3.100	3.840
8	Chris Flexen	SEA	0.977	3.610	3.890	0.700	5.220
9	Blake Snell	SDG	0.978	4.200	3.820	2.400	3.930
10	Robbie Ray	TOR	0.979	2.840	3.690	3.900	3.760

Who? Chris Flexen: 14-7, 2nd most wins in AL

2021 SPR Underperformers

rank	Name	Team	SPR	ERA	FIP	WARP	DRA
218	Carlos Carrasco	NYM	1.020	6.040	5.220	0.500	4.770
219	Aaron Nola	PHI	1.020	4.630	3.370	4.300	3.470
220	Riley Smith	ARI	1.022	6.010	4.880	-0.200	5.870
221	Spenser Watkins	BAL	1.022	8.070	6.370	-0.900	6.980
222	David Peterson	NYM	1.023	5.540	4.770	0.600	4.770
223	Johan Oviedo	STL	1.023	4.910	5.270	0.000	5.560
224	Jackson Kowar	KAN	1.026	11.270	6.430	-0.500	7.160
225	Jake Arrieta		1.030	7.390	6.170	0.100	5.460
226	J.A. Happ		1.032	5.790	5.130	-1.800	6.600
227	Dallas Keuchel	CWS	1.036	5.280	5.220	-2.500	6.890

Closing Thoughts

- While the model seemed to do fairly well at predicting the outcome games during the 2021 season, the problem is still challenging using only moneyline data.

Closing Thoughts

- While the model seemed to do fairly well at predicting the outcome games during the 2021 season, the problem is still challenging using only moneyline data.
- The SPR metric, derived from the model seemed fairly consistent with popular pitching metrics, but demonstrated little predictive value accross seasons.

Closing Thoughts

- While the model seemed to do fairly well at predicting the outcome games during the 2021 season, the problem is still challenging using only moneyline data.
- The SPR metric, derived from the model seemed fairly consistent with popular pitching metrics, but demonstrated little predictive value accross seasons.
- With the growing impact of releif pitchers, we might consider fitting a model for runs scored in the first 5 innings and derive an SPR from these blups.

Data Sources and Further Reading

Data:

- https://www.sportsbookreviewsonline.com/scoresoddsarchives/mlb/mlboddsarchives.htm
- https://www.baseballprospectus.com/leaderboards/pitching/

Jonathan Judge with Baseball Prospectus on DRA:
https://www.baseballprospectus.com/news/article/26196/prospectus-feature-dra-an-in-depth-discussion/

Piper Slowinski with Fangraphs on FIP:
https://library.fangraphs.com/pitching/fip/

Tom Verducci with Sports Illustrated on Starting Pitchers in 2014:
https://www.si.com/betting/2020/07/02/gambling-101-major-league-baseball-betting